REDACT: PII Redaction and Privacy Protection with Ollama Integration

A PROJECT REPORT

Submitted by

MOHAMAD AASIF J (727721EUCS074)

MADHU PRIYA R (727721EUCS064)

MRINALINI K (727721EUCS082)

in partial fulfilment for the award of the degree

of

BACHELOR OF ENGINEERING

IN

COMPUTER SCIENCE AND ENGINEERING

SRI KRISHNA COLLEGE OF ENGINEERING AND TECHNOLOGY

An Autonomous Institution | Approved by AICTE | Affiliated to Anna University | Accredited by NAAC with A++ Grade

Kuniamuthur, Coimbatore – 641008.

APRIL 2025

SRI KRISHNA COLLEGE OF ENGINEERING AND TECHNOLOGY
An Autonomous Institution | Approved by AICTE | Affiliated to Anna University | Accredited by NAAC with A++ Grade
Kuniamuthur, Coimbatore – 641008
Phone : (0422)-2678001 (7 Lines) | Email : info@skcet.ac.in | Website : www.skcet.ac.in

SUSTAINABLE DEVELOPMENT GOALS

The Sustainable Development Goals are a collection of 17 global goals designed to blue print to achieve a better and more sustainable future for all. The SDGs, set in 2015 by the United Nations General Assembly and intended to be achieved by the year 2030, In 2015, 195 nations agreed as a blue print that they can change the world for the better. The project is based on one of the 17 goals.

Questions	Answers
Which SDGs does the project directly address?	The project addresses SDG 16 (Peace, Justice, and Strong Institutions) by ensuring privacy and data protection.
What strategies or actions are being implemented to achieve these goals?	The project integrates AI-powered redaction tools, compliance frameworks, and privacy-enhancing technologies.
How is progress measured and reported in relation to the SDGs?	Progress is tracked through redaction accuracy, compliance adherence, and user feed
How were these goals identified as relevant to the project's objectives?	The goals align with the need for secure data handling, regulatory compliance, and ethical AI implementation.
Are there any partnerships or collaborations in place to enhance this impact?	The project collaborates with legal experts, AI researchers, and data protection organizations.

SRI KRISHNA COLLEGE OF ENGINEERING AND TECHNOLOGY
An Autonomous Institution | Approved by AICTE | Affiliated to Anna University | Accredited by NAAC with A++ Grade
Kuniamuthur, Coimbatore – 641008
Phone : (0422)-2678001 (7 Lines) | Email : info@skcet.ac.in | Website : www.skcet.ac.in

BONAFIDE CERTIFICATE

Certified that this project report **"REDACT: PII Redaction and Privacy Protection with Ollama Integration"** is the bonafide work of **Mr. MOHAMAD AASIF J (727721EUCS074), Ms. MADHU PRIYA R (727721EUCS064), Ms. MRINALINI K (727721EUCS082),** who carried out the project work under my supervision.

SIGNATURE

DR. GRANTY REGINA ELWIN,
HEAD OF DEPARTMENT
Professor
Computer Science and Engineering
Sri Krishna College of Engineering
and Technology Kuniamuthur, Coimbatore

SIGNATURE

Mrs. V. R. AZHAGURAMYAA , M.E.,
SUPERVISOR
Assistant Professor
Computer Science and Engineering
Sri Krishna College of Engineering
and Technology Kuniamuthur, Coimbatore

Submitted for the Project Viva-Voce examination held on _____

INTERNAL EXAMINER

EXTERNAL EXAMINER

ACKNOWLEDGEMENT

We express our sincere thanks to the management and **PORKUMARAN KARANTHARAJ Ph.D.,PEng.,CEng** Principal, Sri Krishna College of Engineering and Technology, Coimbatore for providing us the facilities to carry out this project work.

We are highly indebted to **DR. GRANTY REGINA ELWIN,** Head of Computer Science and Engineering for her continuous evaluation, valuable suggestions and comments given during the course of the project work.

We are thankful to **Ms. A. PRIYA, M.E, (Ph.D),** Project Coordinator**,** Department of Computer Science and Engineering for his continuous evaluation, valuable suggestions and comments given during the course of the project work.

We express our deep sense of gratitude to our guide **Mrs. V. R. AZHAGURAMYAA , M.E.** Professor in the department of Computer science and Engineering for her valuable advice, guidance and support during the course of our project work.

By this, we express our heartfelt sense of gratitude and thanks to our beloved parents, family and friends who have all helped in collecting the resources and materials needed for this project and for their support during the study and implementation of this project.

ABSTRACT

The protection of Personally Identifiable Information (PII) within government documents is an essential aspect of ensuring privacy, data security, and compliance with data protection laws. Traditional methods of manually redacting PII are often slow, labor-intensive, and prone to inaccuracies, especially when dealing with large volumes of documents. This paper introduces a robust and automated PII redaction system designed to address these challenges. The system supports a wide range of document formats, including PDFs and scanned images, making it adaptable for diverse use cases. By leveraging advanced Optical Character Recognition (OCR) technology for efficient text extraction and integrating Large Language Models (LLMs) via Ollama for precise PII identification, the solution guarantees the accurate detection of sensitive information across different languages and document types.

In addition to its core capabilities, the proposed redaction system provides users with the flexibility to review, modify, and customize redaction rules based on their specific needs. This feature allows for greater autonomy in the redaction process, enhancing user control and accuracy. Moreover, the system incorporates cutting-edge document processing techniques, including visual redaction for image-based data, further improving its utility and ease of use. These innovations work together to create a comprehensive solution that not only enhances security but also streamlines the redaction process, making it faster, more efficient, and reliable for handling large-scale document management tasks. Ultimately, this approach ensures greater compliance with data protection regulations while significantly reducing the time and effort required to manage sensitive information within government documents.

TABLE OF CONTENTS

LIST OF FIGURES

CHAPTER 1
INTRODUCTION

1.1 OVERVIEW

The rapid progression of digitalization has revolutionized government and administrative processes, transitioning traditional paper-based systems into electronic formats. These electronic documents are integral to functions such as identification, financial transactions, legal processes, and data exchanges across various sectors. However, many of these documents contain Personally Identifiable Information (PII), which, if exposed or improperly handled, can lead to serious risks such as identity theft, financial fraud, and breaches of privacy.

With the growing volume and complexity of digital documents, ensuring the privacy and security of this sensitive information is more critical than ever. Regulatory frameworks, such as the General Data Protection Regulation (GDPR), Health Insurance Portability and Accountability Act (HIPAA), and California Consumer Privacy Act (CCPA), have been established to safeguard individuals' privacy rights and enforce strict data protection measures.

In this context, it is imperative to develop efficient and accurate methods for redacting PII from digital documents. Traditional redaction techniques often struggle to keep up with the demands of bulk document processing, especially when dealing with different document types, such as scanned images, PDFs, and text-based files.

This has led to the need for more automated, secure, and scalable solutions to handle the redaction of sensitive information. In this paper, we propose a sophisticated redaction framework designed to address these challenges, ensuring PII is securely and accurately removed from a wide array of document formats.

1.2 PII Redaction in Digital Documents

The protection of Personally Identifiable Information (PII) is a critical concern as digital platforms become essential in sectors like government, healthcare, finance, and legal services. Documents containing sensitive data, such as names, Social Security numbers, and addresses, are frequently shared and stored digitally. If exposed, this information can lead to identity theft, financial fraud, and privacy breaches. To prevent such risks, regulations like the General Data Protection Regulation (GDPR), the Health Insurance Portability and Accountability Act (HIPAA), and the California Consumer Privacy Act (CCPA) enforce strict data protection measures, making PII redaction a necessary practice.

Manually redacting PII is a time-consuming and error-prone process, especially when dealing with large volumes of documents. With the increasing reliance on digital records, organizations struggle to keep up with the demands of secure data handling. Errors in manual redaction can lead to significant legal and financial consequences, making it an unsustainable approach for modern enterprises.

Automated redaction solutions offer a more efficient and accurate alternative to manual methods. These systems can process diverse document formats, including PDFs, Word files, scanned images, and handwritten notes. By leveraging Optical Character Recognition (OCR) technology, automated solutions can extract text from image-based documents, enabling accurate identification and removal of PII.

The integration of machine learning and artificial intelligence further enhances the accuracy and adaptability of automated redaction systems. AI-powered models continuously improve by learning from new data, reducing errors and ensuring compliance with evolving regulations. Scalable and efficient, these automated systems help organizations meet the growing demand for secure data handling while ensuring regulatory compliance.

1.3 Redaction Framework

The redaction framework we propose is an advanced, automated system designed to efficiently remove Personally Identifiable Information (PII) from digital documents. As organizations increasingly rely on electronic records, ensuring the privacy and security of sensitive data has become crucial. Our framework integrates Optical Character Recognition (OCR), machine learning, and artificial intelligence (AI) to accurately detect and redact PII while ensuring compliance with regulations such as GDPR, HIPAA, and CCPA.

This system is designed to handle various document types, including PDFs, scanned images, handwritten notes, and structured text files. Its adaptability makes it suitable for industries such as healthcare, government, legal, and finance, where secure data handling is essential. Whether redacting patient records, legal contracts, or financial statements, the framework ensures efficiency and accuracy while reducing the risk of manual errors.

At its core, the framework uses OCR to extract text from image-based and handwritten documents, allowing automated identification of sensitive information. Machine learning continuously improves detection accuracy, reducing false positives and negatives. AI-driven algorithms further enhance adaptability, ensuring precise redaction across different document formats.

Built for scalability and ease of use, the system can process large volumes of documents efficiently. It seamlessly integrates into existing workflows, reducing manual effort while maintaining high security and compliance standards. This framework offers a reliable and efficient solution for organizations seeking to protect sensitive data while streamlining document processing.

1.3.1 APPLICATIONS

- One of the primary applications of our redaction framework lies in enhancing identity authentication and access control systems. The ability to securely redact and protect Personally Identifiable Information (PII) during identity verification processes is crucial across various high-security environments. In settings like government facilities, financial institutions, and military operations, safeguarding sensitive data such as fingerprints, social security numbers, and addresses is paramount. The framework ensures that any stored or processed documents containing PII are appropriately redacted, preventing unauthorized access to this information while supporting secure identity verification methods, including biometric data and smart card technology. This application ensures compliance with strict regulatory requirements and enhances overall security.

- In the realm of law enforcement, our redaction framework is particularly valuable for protecting sensitive case details, witness information, and criminal records while maintaining investigative integrity. For police departments, protecting the privacy of victims, suspects, and law enforcement officers is essential in preserving justice and public safety. The framework can be used to redact PII from police reports, investigation files, and evidence documentation, preventing unauthorized individuals from accessing sensitive data during the course of investigations. Additionally, the system can be used to secure fingerprint and biometric data collected from crime scenes, ensuring that only authorized personnel have access to this data. This capability enhances privacy protection while supporting law enforcement efforts to maintain accurate, confidential records for criminal investigations and court proceedings.

- In the realm of law enforcement, our redaction framework is particularly valuable for protecting sensitive case details, witness information, and criminal records while maintaining investigative integrity. For police departments, protecting the privacy of victims, suspects, and law enforcement officers is essential in preserving justice and public safety. The framework can be used to redact PII from police reports,

investigation files, and evidence documentation, preventing unauthorized individuals from accessing sensitive data during the course of investigations. Additionally, the system can be used to secure fingerprint and biometric data collected from crime scenes, ensuring that only authorized personnel have access to this data. This capability enhances privacy protection while supporting law enforcement efforts to maintain accurate, confidential records for criminal investigations and court proceedings.

- In the financial industry, the redaction framework is particularly valuable for securing sensitive customer and transaction data. Financial institutions handle vast amounts of PII, such as bank account numbers, credit card information, and personal financial histories, all of which must be protected to prevent fraud and identity theft. By utilizing automated redaction, banks and credit card companies can ensure that sensitive customer data is properly secured during routine operations, including data storage, transaction processing, and customer service interactions. Furthermore, the system helps financial organizations maintain compliance with privacy regulations like the General Data Protection Regulation (GDPR) and the California Consumer Privacy Act (CCPA). This application helps safeguard customers' personal and financial information, ensuring a secure, fraud-resistant environment for online banking, transactions, and financial operations.

1.3.2 CHALLENGES

- One of the key challenges in digital redaction is achieving high accuracy when detecting PII, especially in non-textual formats such as images, handwritten documents, or poorly scanned files. Unlike clean, text-based documents that can be easily parsed, images and handwritten content pose significant hurdles for OCR (Optical Character Recognition) systems. OCR technology may struggle with low-quality scans, distorted text, or complex layouts, leading to missed or incomplete redactions. The challenge is ensuring all sensitive data is identified and securely redacted, even in documents with irregular formats, faded text, or illegible handwriting.

- With global organizations handling documents in different languages, ensuring accurate PII detection becomes more complicated. Different countries have unique date formats, address structures, and naming conventions. For example, dates written as DD/MM/YYYY vs. MM/DD/YYYY or names with varying structures can pose challenges for redaction systems. A redaction system must detect and process PII across multiple languages while accounting for regional and cultural differences. This requires advanced language models and natural language processing capabilities to recognize sensitive data accurately in diverse contexts.

- Documents come in many types and formats, including multi-page PDFs, scanned handwritten notes, and image files. These formats may contain PII in various forms, such as embedded images, tables, or unstructured text, making them difficult to redact with a one-size-fits-all approach. The system must identify and redact PII from complex documents, such as multi-column layouts or images with embedded text, while maintaining the integrity of the document's structure. This flexibility is essential for industries like healthcare, legal, and financial services.

1.3.3 FEATURES

- The redaction framework integrates Optical Character Recognition (OCR) technology to extract text from image-based documents, PDFs, and scanned files. This feature allows the system to accurately process text even from non-text formats like images and handwritten documents, enabling effective PII detection and redaction across a variety of document types.

- Using advanced machine learning algorithms, the framework can intelligently detect Personally Identifiable Information (PII) based on predefined criteria. The system continuously improves its accuracy in identifying sensitive data such as names, social security numbers, and financial information, even within complex, unstructured, or poorly formatted documents.

- The redaction framework ensures that sensitive information is not only removed from the text but also visually concealed in image-based data such as scanned forms, photos, and handwritten documents. This added layer of protection guarantees that PII cannot be accessed or exposed, even in image-heavy or graphical data.

- Users can create and apply customizable redaction rules tailored to their organization's privacy and compliance needs. Whether it's for specific industry regulations or company policies, the flexibility of the redaction system allows users to define rules to meet diverse data protection requirements, ensuring proper customization across different use cases.

- The framework provides a user-awareness feature that allows individuals to review and adjust redactions before finalizing the documents. This ensures complete control over the redaction process, providing organizations with the ability to verify and confirm the accuracy of redactions, reducing the risk of errors or oversights.

- The system is designed to be scalable, making it capable of processing large volumes of documents quickly without compromising accuracy or security. Whether handling a few documents or thousands, the framework ensures fast and efficient processing, making it ideal for organizations with high data processing demands.

- The redaction framework supports a variety of document types, including multi-page PDFs, scanned documents, handwritten forms, and images. This versatility ensures that PII can be accurately detected and redacted across all formats, making the solution adaptable to different industries and document management systems.

- The system offers real-time feedback during the redaction process, allowing users to see the effects of their redactions immediately. This helps prevent mistakes and ensures that all necessary PII is accurately removed or concealed before documents are finalized, ensuring transparency and reducing the chances of errors.

- The redaction framework is built to comply with various data protection regulations, such as GDPR, HIPAA, and CCPA. By ensuring that sensitive information is accurately redacted and securely managed, the framework helps organizations meet industry-specific legal requirements and maintain compliance with privacy laws.

CHAPTER 2

LITERATURE SURVEY

Since the advent of e-government services, security of Personally Identifiable Information (PII) is a serious issue now. The conventional redaction techniques—rule-based or manual ones mainly—are not effective in processing unstructured text, handwritten information, and multilingual information, thus becoming useless for batch applications [10][3].

Redaction technology advancements in AI-based technology have provided higher accuracy and automation in the last few years. Machine learning algorithms have also been explored for PII redaction using semi-automated software [3][13]. The majority of these approaches, however, remain template-based or struggle with diverse document structures [3][15]. Large Language Models (LLMs) such as GPT-4o and Llama 3 have been shown to successfully extract and anonymize sensitive data in free-text documents [5][6]. Researchers have also explored privacy-focused LLMs such as Ollama for the safe local processing of data [4]. LLM- based de-identification of sensitive information has also been explored in other applications such as student records and medical reports [14][5].

OCR technology is also used to play a critical part in data extraction from images and scanned documents. The Tesseract OCR engine is utilized extensively for continuous improvement of performance and accuracy in extracting text [7][8][9]. Various studies have compared various PDF parsing tools and have proposed that the need is for quick data extraction from various kinds of documents [15][16]. Combination of OCR with sophisticated AI models enables more effective PII detection, even from sophisticated document structures.

In order to fulfill these requirements, our solution utilizes Ollama (Llama 3.3 70B Versatile) for PII detection as well as PyMuPDF for PDF processing and Tesseract OCR for extracting text from images and scanned documents. Our combined solution provides a privacy-friendly, precise, and scalable redaction solution.

CHAPTER 3
PROBLEM DEFINITION

3.1 OVERVIEW

The increasing reliance on digital services in government and private sectors has led to the collection and processing of vast amounts of Personally Identifiable Information (PII). As the use of electronic documents such as PDFs, scanned documents, and images grows, ensuring the security and privacy of sensitive data becomes a critical challenge. Traditional methods of redacting or anonymizing PII, such as rule-based or manual approaches, are not suitable for modern, unstructured data formats and complex document types. The task of extracting and processing sensitive data from these diverse sources while maintaining accuracy and privacy remains a significant hurdle.

Moreover, many existing solutions struggle with scalability and are not capable of efficiently handling large volumes of data across multiple formats. The need for a robust, automated, and secure method for redacting PII in both text-based and image-based documents is more pressing than ever. Without such a system, organizations risk exposing sensitive information, which could lead to privacy breaches, regulatory violations, and potential cyberattacks. Therefore, there is an urgent need for a comprehensive solution that combines advanced technologies such as Optical Character Recognition (OCR), Natural Language Processing (NLP), and image processing to provide a secure, efficient, and scalable method for PII redaction across a wide range of document formats..

CHAPTER 4

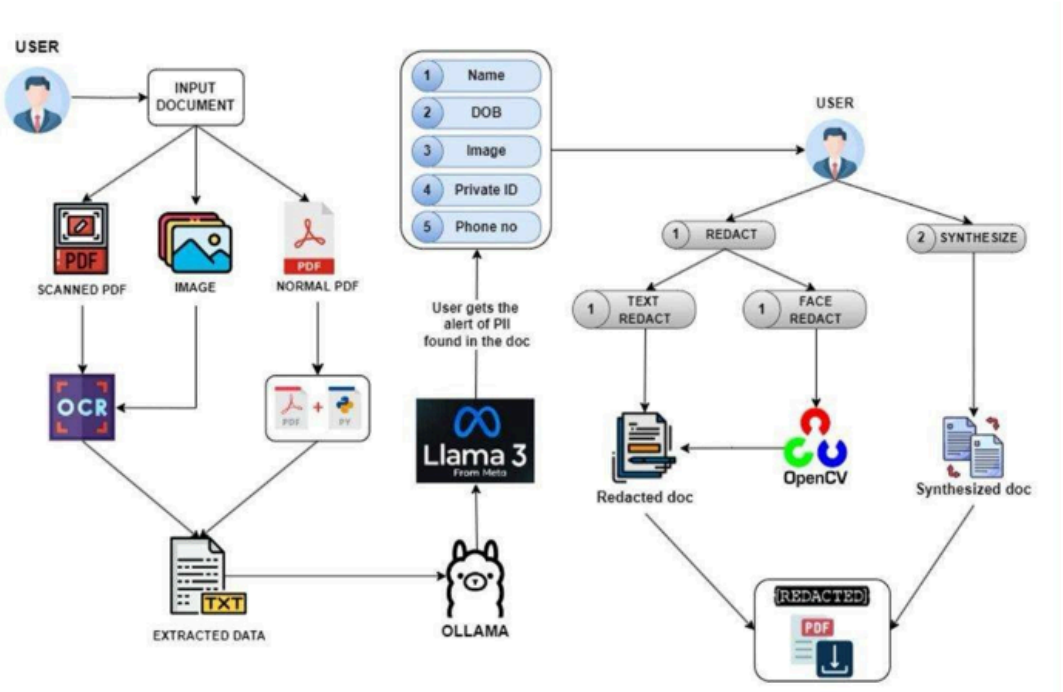
PROPOSED SOLUTION

4.1 OVERVIEW

The **Proposed Solution** integrates a secure and scalable approach to processing sensitive data from various document types, such as PDFs, scanned PDFs, and images. By combining multiple advanced technologies, it ensures high accuracy, privacy, and security throughout the entire workflow. The solution handles the extraction, identification, redaction, and final output of sensitive information while complying with data protection regulations such as **GDPR**, **HIPAA**, and **CCPA**.

At the core of the solution, **PyMuPDF** is used for structured PDFs, while **Tesseract OCR** is employed to process unstructured data from scanned PDFs and images. The system also incorporates **open-sourced large language models (LLMs)** hosted via **Ollama's API** to accurately identify **Personally Identifiable Information (PII)**. This approach ensures that PII is efficiently redacted and anonymized, safeguarding privacy and confidentiality.

The solution is designed for scalability, making it ideal for organizations processing large volumes of sensitive documents. It is adaptable to various use cases, with robust redaction capabilities for both text and visual data.



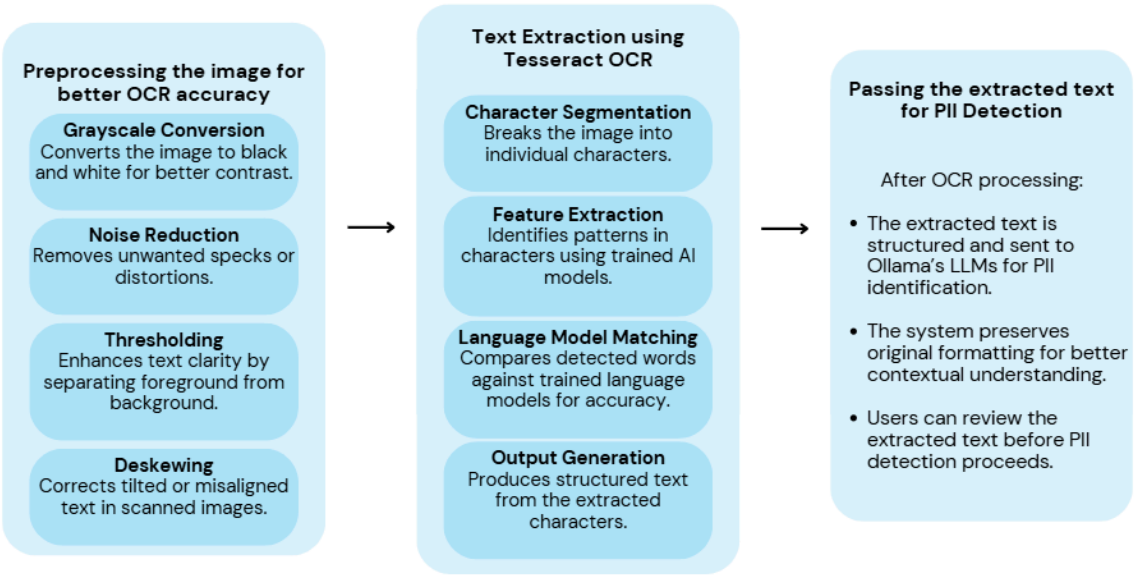
4.2 DATA EXTRACTION

The **Data Extraction** phase of the solution is where the system extracts content from various document formats, including structured and unstructured documents.

- **PyMuPDF**: For structured PDFs, **PyMuPDF** is used to efficiently parse and extract text, images, and other content from the document while preserving its original structure. This is ideal for documents that contain formatted tables, graphs, and complex layouts.

- **Tesseract OCR**: For scanned PDFs and image-based documents, **Tesseract OCR** is used to extract text from images. Tesseract converts scanned images and documents into machine-readable formats, enabling the system to process even poorly scanned or low-quality documents.

The combination of these two tools ensures that the system can handle both structured and unstructured data with high accuracy.



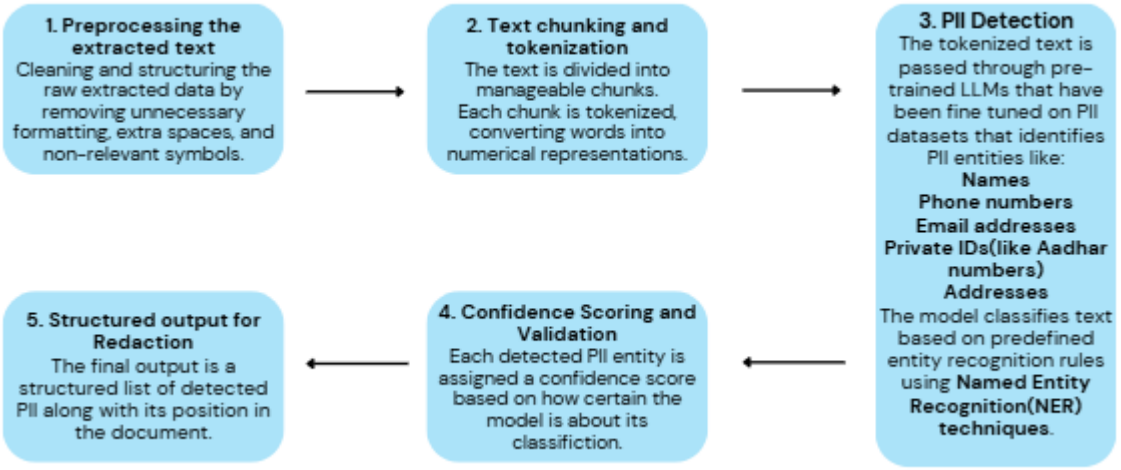
4.3 PII IDENTIFICATION

Once the data is extracted, the next step is **PII Identification**, where **open-sourced large language models (LLMs)** come into play.

- **LLMs for PII Detection**: After the data has been extracted, **LLMs** are employed to scan the content and identify **Personally Identifiable Information (PII)**. These models are designed to detect various types of PII, including names, addresses, email addresses, phone numbers, social security numbers, and more. By analyzing the context in which these identifiers appear, LLMs ensure that sensitive data is correctly identified and flagged.

- **Contextual Understanding**: The system uses LLMs' contextual understanding to ensure accurate identification of PII, even in cases where it is not explicitly stated, minimizing the chances of false positives.

By hosting these models securely within the system, the solution maintains strict data privacy by keeping all processing within the organization's infrastructure.



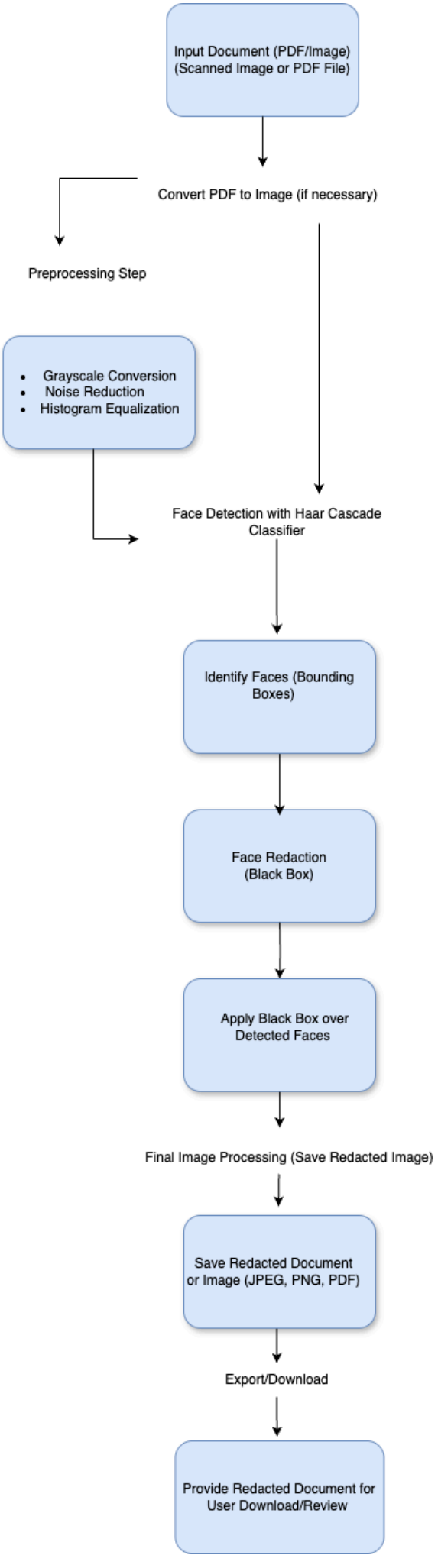
4.4 REDACTION AND ANONYMIZATION

The **Redaction and Anonymization** phase is where sensitive data is securely obscured to ensure privacy.

- **Textual Redaction**: For textual PII, **PyMuPDF** places black boxes over the identified sensitive data, ensuring it cannot be recovered. This redaction is permanent, making the information unreadable and secure.

- **Visual Redaction**: For visual data, **OpenCV** is employed to detect facial features in images and scanned documents. Once identified, facial features are anonymized by placing black boxes over them, ensuring that visual identifiers are completely obscured.

- **Data Synthesis**: To further enhance security, the system synthesizes redacted PII by replacing it with unrecognizable, synthetic data. This added layer of protection ensures that even if the redacted data is recovered, it cannot be linked to any real individual.


Input Document (PDF/Image)
(Scanned Image or PDF File)
Convert PDF to Image (if necessary)
Preprocessing Step
Grayscale Conversion
Noise Reduction
Histogram Equalization
Face Detection with Haar Cascade Classifier
Identify Faces (Bounding Boxes)
Face Redaction (Black Box)
Apply Black Box over Detected Faces
Final Image Processing (Save Redacted Image)
Save Redacted Document or Image (JPEG, PNG, PDF)
Export/Download
Provide Redacted Document for User Download/Review

4.5 SECURITY AND PRIVACY

The **Security and Privacy** aspect is central to the design of the system, ensuring that sensitive data is protected throughout the entire process.

- **Local Processing**: All data is processed locally within the system, which ensures that sensitive information is not transmitted to third-party servers. This approach minimizes the risk of data breaches and maintains data control within the organization.

- **Synthesis and Obfuscation**: In addition to traditional redaction, the system also synthesizes redacted data to make it unrecognizable. This provides an extra layer of security by preventing attackers from attempting to reverse-engineer the redacted content.

- **Compliance with Regulations**: The system is built to comply with contemporary data protection regulations, such as **GDPR**, **HIPAA**, and **CCPA**, ensuring that all sensitive data processing meets the highest standards of privacy and security.

By integrating these features, the solution ensures that sensitive data is securely processed and anonymized, preventing unauthorized access and maintaining privacy.

CHAPTER 5

RESULT

5.1 USER UPLOAD AND DOCUMENT PROCESSING

The system can process user uploads of both single and multiple files efficiently, supporting structured PDFs, scanned PDFs, and image file types like PNG, JPEG, and TIFF. When a single file is uploaded, it is processed and redacted. In the case of multiple file uploads, each document undergoes consistent extraction, redaction, and security measures before being compiled into a ZIP file for download. For structured PDFs, PyMuPDF is used to maintain the document's layout and structure while extracting the necessary text. Scanned PDFs and images are processed with Tesseract OCR, which accurately extracts text from unstructured content, ensuring that text extraction is highly precise regardless of the document type.

5.2 PII DETECTION, REDACTION, AND ANONYMIZATION

Once the data extraction is complete, open-source Large Language Models (LLMs) hosted via Ollama's API are employed to identify Personally Identifiable Information (PII) like names, addresses, phone numbers, and other sensitive details in the extracted text. The LLMs analyze the content contextually to ensure that all potential PII is identified promptly and accurately. To redact sensitive information, PyMuPDF is used to stamp black boxes over the PII, ensuring that the redacted content is irrecoverable, thus maintaining document integrity. OpenCV is also used to detect facial landmarks in scanned documents and images, applying black boxes to mask faces for visual privacy. Finally, redacted data is anonymized to prevent re-identification, offering an added layer of security to ensure the system complies with privacy laws and protects against cyber threats.

5.3 SECURITY, SCALABILITY, AND WORKFLOW INTEGRATION

The software operates in a secure, isolated environment, ensuring that all data is processed locally without being transmitted to external servers, allowing the user to maintain control over their sensitive information. This setup helps protect against third-party API service and cloud storage vulnerabilities, offering robust data protection. Once redacted, the data is anonymized again to prevent identification, meeting the highest standards of data protection and ensuring compliance with privacy regulations. The system is designed for scalability, with the ability to handle large volumes of documents of varying complexity. This minimizes human intervention and boosts business efficiency, while a simple and intuitive user interface makes it easy for users to inspect and validate redacted information.

CHAPTER 6

TECHNOLOGIES

6.1 PYTHON

Python is the core programming language for our system due to its simplicity, versatility, and wide range of libraries available for document processing and machine learning. In this project, Python is used to handle file uploads, extract text from PDFs and images, apply redaction, and implement secure processing and anonymization workflows. With libraries such as PyMuPDF for structured PDF handling, Tesseract OCR for optical character recognition, and OpenCV for image processing, Python is the ideal choice for performing complex document extraction and redaction tasks. Python's efficiency and ease of integration with other technologies ensure seamless system performance and flexibility.

6.2 FASTAPI

FastAPI is used to build the backend of the application, ensuring fast and efficient handling of HTTP requests. This modern web framework allows us to create a RESTful API that facilitates seamless communication between the frontend (built with React) and the backend processes. FastAPI's asynchronous capabilities make it perfect for handling multiple file uploads and processing tasks simultaneously, ensuring low latency and high throughput. The API is designed to accept file uploads, process documents (e.g., extraction, redaction), and return results to the user, all while maintaining performance and scalability. FastAPI also provides automatic data validation and documentation through OpenAPI, making it easy to integrate and scale the application.

6.3 REACT JS

React is used to build the frontend of the application, ensuring a dynamic and responsive user interface. React's component-based architecture makes it easier to develop and maintain complex UIs. The frontend is responsible for allowing users to upload single or multiple files, inspect redacted information, and download processed documents. With React's state management and hooks, the UI can interact seamlessly with the FastAPI backend, handling file uploads and displaying processing results in real time. React's flexibility allows for smooth user interactions and ensures that the system can easily scale with new features or improvements in the future.

6.4 OPENCV

OpenCV (Open Source Computer Vision Library) is used in the project for facial recognition and image processing. OpenCV helps detect facial landmarks in images or scanned documents, ensuring that sensitive visual data, such as faces in photographs, is appropriately redacted. The tool automatically identifies facial features and applies redaction (e.g., black boxes) to mask them, ensuring visual privacy. This ensures that even documents containing photographs are handled securely and that sensitive visual information is protected, adhering to privacy regulations and data security standards.

6.5 Tesseract OCR

Tesseract OCR is a powerful optical character recognition tool that plays a crucial role in extracting text from scanned PDFs and image files (e.g., PNG, JPEG, TIFF). When documents are unstructured or in image format, Tesseract OCR enables the system to recognize and extract text with high accuracy. By converting images to machine-readable text, Tesseract ensures that even unstructured content can be processed and redacted for PII detection. This functionality is essential for documents that do not contain structured data but still contain sensitive information that needs to be redacted or anonymized.

6.6 Ollama

Ollama provides the infrastructure for running open-source Large Language Models (LLMs) that are employed to identify Personally Identifiable Information (PII) within the extracted document data. These models perform contextual analysis to identify sensitive data such as names, addresses, phone numbers, and other PII types. Ollama's hosted API allows the system to process large volumes of text while maintaining high accuracy in PII detection, ensuring that all sensitive information is detected and flagged for redaction. By utilizing Ollama, the system can leverage the power of advanced machine learning models to enhance data processing capabilities and ensure robust PII detection without relying on external third-party services.

CHAPTER 7

CONCLUSION AND FUTURE WORKS

7.1 CONCLUSION

In all, the proposed system provides a safe and solid platform for redaction, detection, and anonymization of personally identifiable information (PII) for most document types. Through the application of advanced tools such as PyMuPDF, Tesseract OCR, OpenCV, and open-source large language models (LLMs), it carries out accurate data extraction, effective processing, and uncompromising redaction of confidential information. Utilization of internal PII detection model hosting circumvents foreign platforms, hence evading data breach risks. Additionally, the system provides irretrievable redaction of confidential information, satisfying very high data protection standards. Through its modularized design, scalability and flexibility ensure the system efficiently performs with various applications in real- world usage.

While technical excellence in the system is backed by passion for information protection and ease of use, automation and seamless workflows are incorporated into enterprise networks with ease. High accuracy and performance are guaranteed under various levels of demands. Using leading-edge technology and a privacy-first design, the solution is able to overcome data security challenges in contemporary times. Future development can possibly enhance processing efficiency and expand the scope of redaction, but the system needs to be a capable and forward- looking solution for business domains dealing with confidential information on a large scale.

7.2 FUTURE ENHANCEMENT

Some directions in which the application can be enhanced for performance, scalability, and usability are possible. Priority-level redaction support is a major one among them. By this feature, users will have the ability to assign various priorities to redact confidential information based on the level of confidence with which they treat it as confidential. For instance, if one needs to redact information such as on an Aadhar card, one might prioritize name as low priority, address as medium, and Aadhar number as high priority. Integrating this would help the system cover more cases of usage and thus provide flexibility for the user, providing him more control of how he is to redact things.

One other area which really needs improving is the Faker module. The Faker module is utilized to replace redacted data with pseudo-random but realistic existing faker providers' Personally Identifiable Information (PII). This functionality can be further extended by improving the region-based PII synthesizing feature with more sophisticated capability to create localizable data in order to better meet the special needs and regulatory environments of a specific geographic area. This would improve usability for downstream processing while still allowing efficient privacy protection.

These enhancements would all contribute towards making the application more flexible, secure, and customizable to fit various user requirements and various region requirements.

APPENDIX

A.1 SOURCE CODE:

main.py:

```
from fastapi import FastAPI, File, UploadFile, Form
from fastapi.responses import FileResponse, HTMLResponse
from fastapi.templating import Jinja2Templates
import os
import tempfile
import logging
import pandas as pd
import json
import zipfile
import re
from io import BytesIO
from pathlib import Path
from typing import List
from core import read_pdf, search_replace_in_pdf, read_image, search_replace_in_image
from pydantic import BaseModel, ValidationError

app = FastAPI()

logging.basicConfig(level=logging.INFO)

templates = Jinja2Templates(directory="templates")

def is_pdf_or_image(file):
    file_ext = os.path.splitext(file.filename)[1].lower()
    if file_ext == ".pdf":
```

```
        return "pdf"
    elif file_ext in [".png", ".jpg", ".jpeg"]:
        return "image"
    return None

class IdentifierType(str, Enum):
    EMAIL = "Email"
    GOVERNMENT_ID = "Government ID Number"
    NAME = "Name"
    PHONE_NUMBER = "Phone number"
    ADDRESS = "Address"
    UNKNOWN = "Unknown"
    ENROLMENT_NO = "Enrolment No."
    FATHERS_NAME = "Father Name"
    SURNAME = "Surname"
    VID = "VID"
    DATE_OF_BIRTH = "Date of Birth"
    PLACE_OF_BIRTH = "Place of Birth"
    DATE_OF_EXPIRY = "Date of Expiry"
    DATE = "Date"
    AADHAAR_ISSUE_DATE = "Aadhaar Issue Date"
    PIN_CODE = "PIN Code"
    Place_of_Issue = "Place of Issue"

class Identifier(BaseModel):
    objValue: str
    objType: IdentifierType

def extract_entities(text: str):
    identifiers = []
```

```
    identifiers.append(Identifier(objValue="John Doe", objType=IdentifierType.NAME))
    identifiers.append(Identifier(objValue="9876543210",
objType=IdentifierType.PHONE_NUMBER))
    return identifiers


def read_file(file, file_type):
    if file_type is None:
        raise ValueError("Invalid file type")

    with tempfile.NamedTemporaryFile(delete=False) as temp_file:
        temp_file.write(file.file.read())
        temp_file_path = temp_file.name

    if file_type == "pdf":
        return read_pdf(temp_file_path)
    elif file_type == "image":
        return read_image(temp_file_path)
    else:
        raise ValueError(f"Unsupported file type: {file_type}")

def search_replace(file, words: list[str], file_name: str, remove_picture: bool):
    red_file_name, red_file_ext = file_name.split(".", 1)
    red_file_name = f"{red_file_name}_redacted.{red_file_ext}"
    file_type = is_pdf_or_image(file)

    with tempfile.NamedTemporaryFile(delete=False) as temp_file:
        temp_file.write(file.file.read())
        temp_file_path = temp_file.name

    redacted_file_path = None
```

```
    match file_type:
        case "pdf":
            redacted_file_path = search_replace_in_pdf(temp_file_path, words, remove_picture,
red_file_name)
        case "image":
            redacted_file_path = search_replace_in_image(temp_file_path, words,
remove_picture, red_file_name)
        case _:
            raise ValueError("Invalid file type")

    if redacted_file_path is None:
        raise ValueError("Failed to obtain a valid redacted file path.")

    return redacted_file_path

def zip_redacted_files(file_paths: list[str]) -> str:
    zip_filename = "redacted_files.zip"
    with zipfile.ZipFile(zip_filename, 'w') as zipf:
        for file_path in file_paths:
            zipf.write(file_path, os.path.basename(file_path))
    return zip_filename

@app.get("/", response_class=HTMLResponse)
async def read_root(request):
    return templates.TemplateResponse("index.html", {"request": request})

@app.post("/uploadfile/")
async def upload_file(file: UploadFile = File(...)):
    file_type = is_pdf_or_image(file)

    if not file_type:
```

```
        return {"error": "Invalid file type"}

    with tempfile.NamedTemporaryFile(delete=False) as temp_file:
        temp_file.write(await file.read())
        temp_file_path = temp_file.name

    extracted_data = extract_entities(temp_file_path)

    return {"extracted_data": [data.dict() for data in extracted_data]}

@app.post("/redact/")
async def redact_file(file: UploadFile = File(...), data_to_redact: List[str] = Form(...)):
    file_type = is_pdf_or_image(file)

    if not file_type:
        return {"error": "Invalid file type"}

    with tempfile.NamedTemporaryFile(delete=False) as temp_file:
        temp_file.write(await file.read())
        temp_file_path = temp_file.name

    redacted_file_path = search_replace(temp_file_path, data_to_redact, file.filename,
remove_picture=True)

    return FileResponse(redacted_file_path, media_type='application/octet-stream',
filename=os.path.basename(redacted_file_path))

if __name__ == "__main__":
    import uvicorn
    uvicorn.run(app, host="0.0.0.0", port=8000)
```

A.2 SCREENSHOTS

Fig.A.1 Input

Fig A.1 The user inputs a file or set of files with PII present

	objValue	objType
0	दीपक	Name
1	कपूर	Name
2	Deepak	Name
3	Kapoor	Name
4	22/06/1983	Date of Birth
5	5939	Government ID Number
6	7553	Government ID Number
7	9390	Government ID Number

Fig.A.2 Successful Extraction and Identification ofPII in document in out UI

Then the user will be able to choose the type of PII to redact from our User Interface

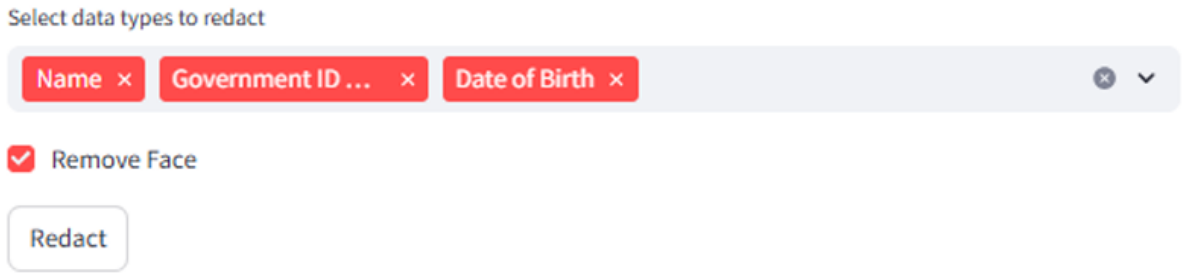


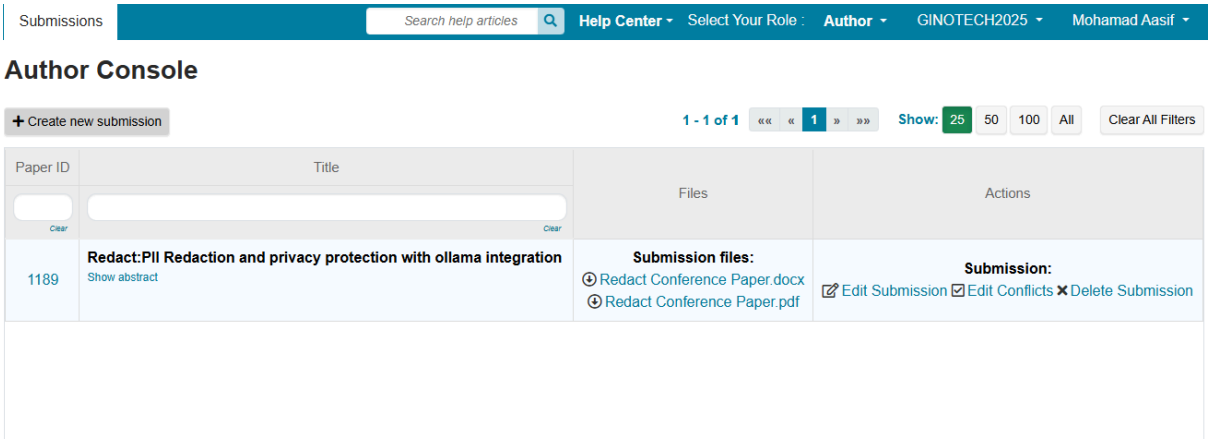
Fig.A.3 Dropdown with all the type of PII identified and if user wants to remove face if found

If there is no face present only the selected PII types will be redacted and if multiple files where selected then the user will be able to download a zip file of the REDACTED files.

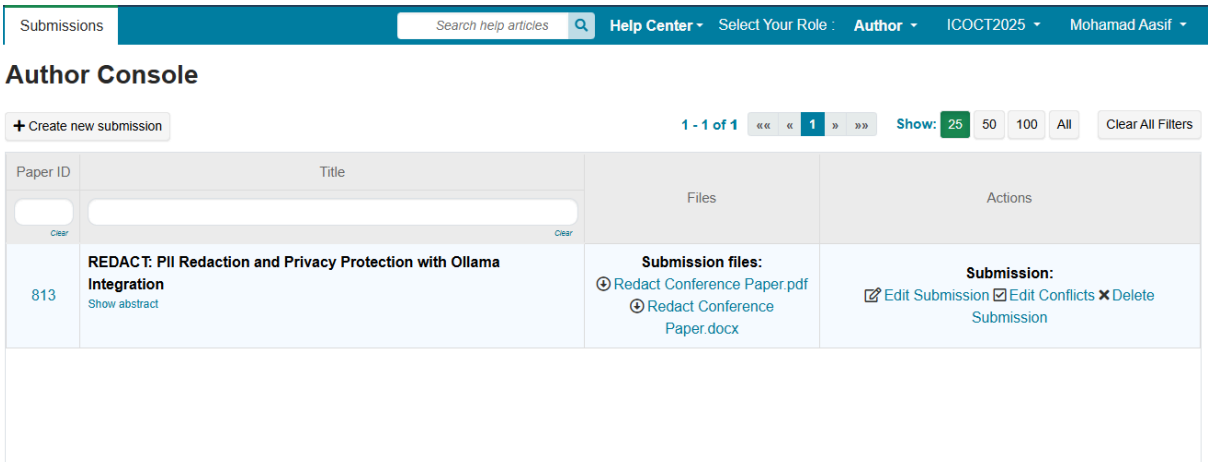
Fig.A.4 Output downloadable REDACTED file with the selected PII types and face is redacted

The PII is detected by Ollama and redacted by OCR for scanned pdf and images and PyMuPDF for pdf documents.

A.3 CONFERENCE AND PUBLICATION



IEEE GLOBAL CONFERENCE IN EMERGING TECHNOLOGY 2025 Date : 09th May-11th May 2025 , Venue- Dr. D. Y. Patil Institute of Technology, Pimpri Pune



International Conference On Computing Technologies

Jyothy Institute of Technology

8th International Conference on Computing Methodologies and Communication (ICCMC 2025)
23-25, July 2025
https://icocmc.com/ICCMC-25/
info.iccmc@gmail.com

Acceptance Letter

Details of Accepted Paper

Paper ID - ICCMC-012

Title - REDACT: PII Redaction and Privacy Protection with Ollama Integration

Author(s) - Azhaguramyaa VR, Mohamad Aasif J, Mrinalini K , Madhu Priya R

Dear Author,

Greetings from ICCMC 2025!!

The Organizing Committee is pleased to inform you that the above peer-reviewed & refereed conference paper has been accepted for presentation at the **8th International Conference on Computing Methodologies and Communication (ICCMC 2025)** organized by *Surya Engineering College, Erode, Tamil Nadu, India* on 23-25, July 2025.

ICCMC 2025 will provide an exceptional international forum for sharing knowledge and results in all fields of engineering and technology. It requires quality key experts who provide an opportunity in bringing up innovative ideas. Recent updates in the technology will be a platform for upcoming researchers. This conference will offer you an unforgettable experience in exploring new opportunities.

With Thanks,
Yours' Sincerely

ICCMC

Conference Chair
ICCMC 2025

REFERENCES

[1] Tianyu Yang, Xiaodan Zhu, Iryna Gurevych, Robust utility-Preserving text anonymization based on large language models

[2] Emiliano De Cristofaro, University College London, What is Synthetic Data? The Good, The Bad, and the Ugly

[3] Chad Cumb, Rayid Ghani, A machine learning based system for semi-automatically redacting documents

[4] Rodríguez Quiñones, Adrià, Privacy-Focused LLM for Local Data Processing: Implementing OLLAMA and RAG to Securely Query Personal Files in Closed Environments

[5] Jonas Wihl, Enrike Rosenkranz, Severin Schramm, Cornelius Berberich, Michael Griessmair, Piotr Woźnicki, Francisco Pinto, Sebastian Ziegelmayer, Lisa C. Adams, Keno K. Bressem, Jan S. Kirschke, Claus Zimmer, Benedikt Wiestler, Dennis Hedderich, Su Hwan Kim, Data Extraction from Free-Text Stroke CT Reports Using GPT-4o and Llama- 3.3-70B: The Impact of Annotation Guidelines

[6] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian +470 more authors, The Llama 3 Herd of Models

[7] R. Smith, An Overview of the Tesseract OCR Engine

[8] Sahil Badla, IMPROVING THE EFFICIENCY OF TESSERACT OCR ENGINE

[9] Tao Ma, Min Yue, Chao Yuan, Haibo Yuan, File Text Recognition and Management System Based on Tesseract- OCR

[10] PM Schwartz, DJ Solove, The PII problem: Privacy and a new concept of personally identifiable information

[11] Y Yao, J Duan, K Xu, Y Cai, Z Sun, Y Zhang, A survey on large language model (LLM) security and privacy: The good, the bad, and the ugly

[12] Jingfeng Yang, Hongye Jin, Ruixiang Tang, Xiaotian Han, Qizhang Feng, Haoming Jiang, Shaochen Zhong, Bing Yin, Xia Hu, Harnessing the Power of LLMs in Practice: A Survey on ChatGPT and Beyond

[13] Shobin Joyakin, iMask—An Artificial Intelligence Based Redaction Engine

[14] Shreya Singhal, Andres Felipe Zambrano, Maciej Pankiewicz, Xiner Liu, Chelsea Porter, Ryan S. Baker, De- Identifying Student Personally Identifying Information with GPT-4

[15] Narayan S. Adhikari, Shradha Agarwal, A Comparative Study of PDF Parsing Tools Across Diverse Document Categories

[16] Rohaan Nadeem, Tahir Iqbal, Noor Fatima, Junaid Altaf, Asma Irshad, Asif Farooq, Extraction of User-Defined Information from PDF